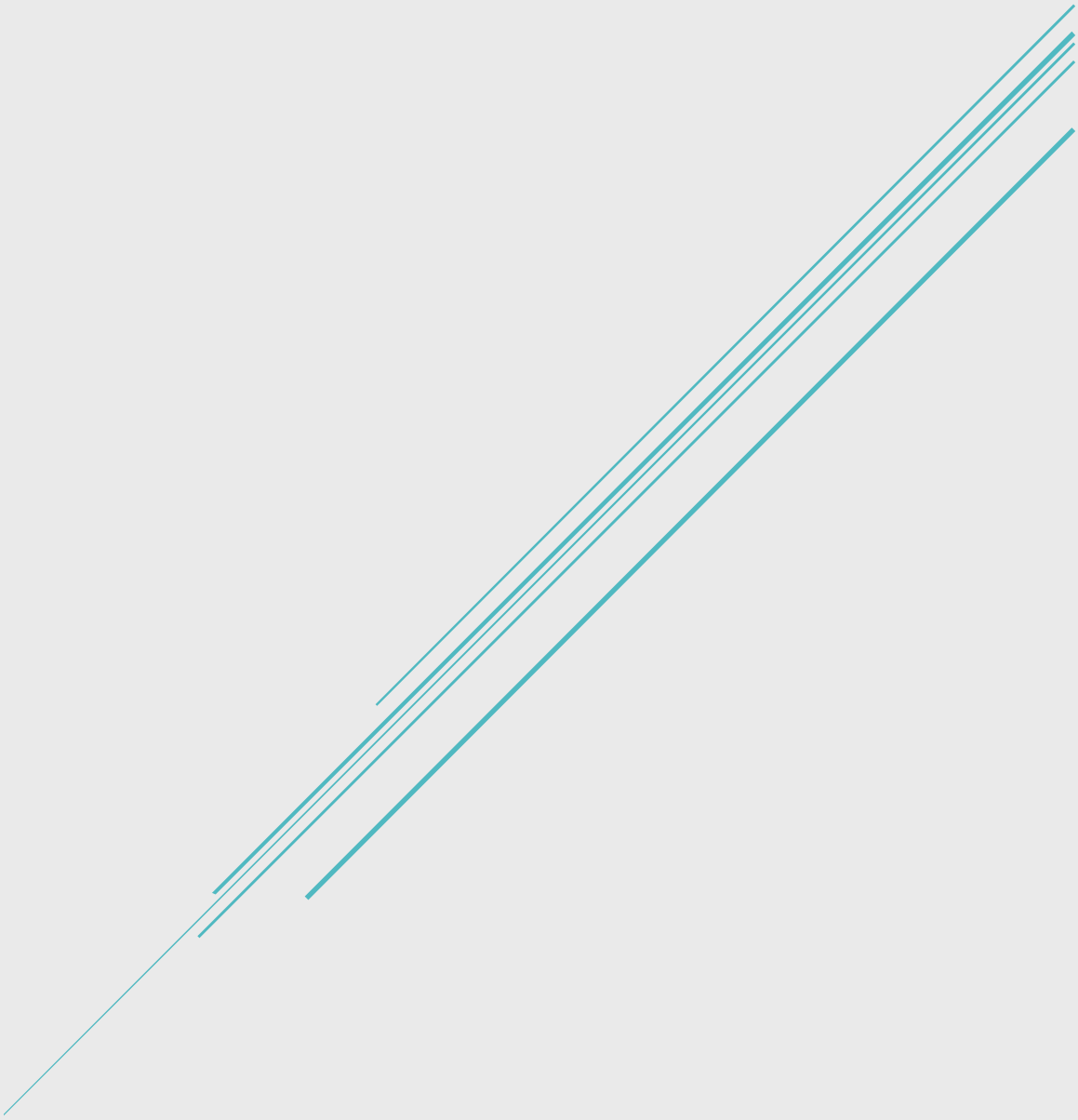
PROYECTO 1: ANALISIS DE RESEÑAS DE SITIOS TURISTICOS

Grupo 1



Integrantes

Medina Martínez, Ana Sofia. Pérez Castilla, Carlos Mario. Castellanos Bonilla, Juan Diego

CONTENIDO

INTRODUCCIÓN

El proyecto es de analítica reseñas de sitios turísticos, usando información textual (recopilaciones), donde se quiere clasificar las reseñas de los turistas en Colombia De manera que con el modelo en construcción se pueda facilitar la clasificación de reseñas. Para poder hacer un análisis automatizado de opiniones que representan la voz de los turistas sobre sitios de turismo el cual visitaron.

Este modelo podría llegar a tener un beneficio en El Ministerio de Comercio, Industria y Turismo de Colombia, la Asociación Hotelera y Turística de Colombia – COTELCO, cadenas hoteleras de la talla de Hilton, Hoteles Estelar, Holiday Inn y hoteles pequeños ubicados en diferentes municipios de Colombia están interesados en analizar las características de sitios turísticos que los hacen atractivos para turistas locales o de otros países. Pues la ventaja está en que rápidamente se podría llegar a una idea de cuáles son las falencias y cuáles son las problemáticas de su entorno para tomar decisiones sobre cómo se puede abordar este problema en el entorno colombiano y de esta forma poder aplicar estrategias para identificar oportunidades de mejoras que permitan aumentar la popularidad de estos y de esta manera aumentar el turismo.

(10%) ENTENDIMIENTO DEL NEGOCIO Y ENFOQUE ANALÍTICO.

Oportunidad/problema Negocio	usando información textual (recopilaciones), donde se quiere clasificar las reseñas de los diferentes sitios turísticos. Se quiere Construir un modelo que clasifique la información textual.
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático) e incluya las técnicas y algoritmos que propone utilizar.	Teniendo en cuenta que los valores dados eran (información textual y un valor numérico el cual decía la calificación acorde a la reseña) y que el enunciado dice que se debe clasificar el texto. Se decidió implementar técnicas de clasificación supervisada. Con el fin de tener un análisis correcto. se decidió que cada estudiante implementara un algoritmo, dando un total de 3 entre estos esta: SVM, Arboles de decisión (RandomForestRegressor) y Naive Bayes (Multinomial). De igual manera para el manejo de los datos para los modelos se utilizó un pipeline
Organización y rol dentro de ella que se beneficia con la oportunidad definida	La organización que se beneficia con la oportunidad definida: Construcción de un modelo que clasifique la información textual. Son El Ministerio de Comercio, Industria y Turismo de Colombia, la Asociación Hotelera y Turística de Colombia – COTELCO, cadenas hoteleras de la talla de Hilton, Hoteles Estelar, Holiday Inn y hoteles pequeños ubicados en diferentes municipios de Colombia. En cuanto a los roles, teniendo en cuenta que una de las organizaciones que se benefician son el ministerio de comercio e industria, unos de los roles beneficiados serian los siguientes:

	1. Responsables de Políticas y Planificación Turística. 2. Departamento de Promoción y Mercadeo Turístico. 3. la Asociación Hotelera y Turística de Colombia (COTELCO).
Contacto con experto externo al proyecto	En cuanto al contacto con un experto al proyecto se tiene un estudiante de estadística. En el caso del grupo 1 son tres estudiantes: 1. Manuela Insuasti Vargas y su correo de contacto es: m.insuastiv@uniandes.edu.co 2. Sara Hernández y su correo de contacto es: ss.hernandezm1@uniandes.edu.co 3. Lina Arias y su correo de contacto es: lm.ariasp1@uniandes.edu.co

(20%) ENTENDIMIENTO Y PREPARACIÓN DE LOS DATOS.

Entendimiento de datos

- 7875 datos
- 1 columna textual (Review)
- 1 columna categórica (Class)

Completitud

- Todos los datos están completos, es decir ninguna celda esta vacía.

Unicidad

- Acorde al Pandas Profiling Report se encontraron 29 datos duplicados en Review dando un duplicado de 0.4%
- Los valores de Class se repiten los cuales son 1,2,3,4,5

Nulos

- No hay datos nulos

Validez

- Todos los datos son válidos, vale aclarar que se limpiaran los datos repetidos mas adelante.

Limpieza y tratamiento de datos:

Limpieza de datos:

1. Se eliminó los datos duplicados

Preparación de datos

1. **Se eliminaron todos los caracteres no ASCII**
   el motivo de eliminar estos caracteres no ASCII es que no pertenecen al alfabeto latino.
2. **Todo carácter pasó a estar en minúscula**
   esto se realizó para evitar problemas al momento de trabajar con palabras y que sean diferentes.
3. **Se elimino toda puntuación (./,/;/ (/))**
   no aporta mucha información al momento de usar la raíz de la palabra
4. **Se remplazó todo número a palabras con number_to_words**
   como se tienen que pasar todos los valores a numérico para los modelos se remplazan todos los numero a texto para evitar errores.
5. **Se realiza un WordCloud el cual nos permite ver un análisis de las palabras más recurrentes dentro de cada categoría.**
6. **Se cambiaron las palabras vacías (comunes) con stopwords con la librería nltk a espacios.**
   quitar las palabras comunes es importante, ya que estas no aportan nada al texto, ni al análisis.
7. **Con countVectorize y TfidfVectorizer se construyó una matriz que mirara cuantas, si aparece una palabra en especifico**
   Esto se realizo con el fin de poder pasar las palabras a una representación numérica y de esta forma poder clasificarlos.
8. **Se realizo una tokenización permite dividir frases u oraciones en palabras.**
   Esto permite desglosar las palabras correctamente para el posterior análisis. Pero primero, se realiza una corrección de las contracciones que pueden estar presentes en los textos.
9. **Se realiza normalización en el texto:**
   Esto permite realizar una eliminación de prefijos y sufijos, además de realizar una lematización de los verbos.

Técnicas y algoritmos

- SVM (máquinas de vectores de soporte)
- Arboles de decisión(Random Forest Regressor, Random Forest Classifier)

- Naive Bayes (multinomial)

Técnica: Clasificación

(25%) MODELADO Y EVALUACIÓN.
SVM (máquinas de vectores de soporte)

El algoritmo de máquinas de vectores de soporte es un algoritmo de aprendizaje supervisado, el cual se usó para la clasificación binaria y poder separar de la mejor forma posible dos clases diferentes de puntos de datos.

Test Report for Support Vector Machines

	precision	recall	f1-score	support
1	0.42	0.46	0.44	144
2	0.33	0.32	0.32	230
3	0.35	0.40	0.37	296
4	0.39	0.36	0.38	404
5	0.61	0.58	0.59	487
accuracy			0.44	1561
macro avg	0.42	0.42	0.42	1561
weighted avg	0.44	0.44	0.44	1561

```
Precisión del modelo con test : 0.4407431133888533
```

Arboles de decisión

Random Forest Regressor radica en el hecho de que combina múltiples árboles de decisión, lo que mejora la precisión y la capacidad de generalización del modelo.

```
Métricas de rendimiento en el conjunto de entrenamiento:
Mean Squared Error (MSE): 0.9906103139013454
Mean Absolute Error (MAE): 0.7810570147341447
R-squared (R^2): 0.4239095314410035
```

RandomForestClassifier: es un algoritmo de aprendizaje supervisado que se utiliza para la clasificación. Se basa en el concepto de ensamblado, donde se combinan múltiples árboles de decisión para obtener una predicción más precisa y robusta.

```
Precisión del modelo: 0.45996156310057656
```

```
F1 Score del modelo: 0.4368240640945578
```

Naive bayes

Naive Bayes es un algoritmo de aprendizaje automático utilizado principalmente para problemas de clasificación en aprendizaje supervisado. Este algoritmo se basa en el teorema de Bayes y asume independencia condicional entre los atributos o

características.

Train Report for naive bayes multinomial

	precision	recall	f1-score	support
1	0.54	0.23	0.32	144
2	0.38	0.37	0.37	230
3	0.34	0.36	0.35	296
4	0.43	0.43	0.43	404
5	0.60	0.70	0.65	487
accuracy			0.47	1561
macro avg	0.46	0.42	0.42	1561
weighted avg	0.47	0.47	0.46	1561

Precisión del modelo con test : 0.47085201793721976

Mejor técnica: Naive Bayes

Valor: **0.47**

(15%) RESULTADOS.

En cuanto a los resultados obtenidos entre todos los modelos que se realizaron se tiene en cuenta que todos los algoritmos implementados son de clasificación de texto. De igual manera cabe mencionar que en todos estos se realizó la misma limpieza, de igual manera se verifico con: countVectorize y TfidfVectorizer y se quedó con las mejores métricas teniendo en cuenta cómo funciona el algoritmo y seleccionar el mejor entre los dos. Sin embargo, la comparación de la precisión del algoritmo dio como resultado que el modelo Naive Bayes fue el mejor de todos con un total de **0.47**. Esto nos dice que los datos están en mal estado y que el modelo no es bueno para la toma de decisiones, ya que tiene un 0.53% de margen de error. De igual manera al calcular el F1 Score para todos los modelos se pudo observar como el modelo Naive Bayes tuvo un promedio de F1 del 0.47, mientras que los arboles fueron de 0.43 y SVM de 0.44

Una vez se sabia esto se recomienda a la Organización de las naciones unidas y a la UNFPA el uso de los siguientes modelos: SVM ya que tuvo la mejor precisión y un F1 Score de 0.99.

MAPA DE ACTORES RELACIONADO CON UN PRODUCTO DE DATOS CREADO CON EL MODELO ANALÍTICO CONSTRUIDO

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Ministerio de Comercio, Industria y Turismo de Colombia	Patrocinador / Tomador de Decisiones (Cliente)	Obtener insights valiosos sobre las características y percepciones de los turistas en los sitios turísticos del país. Diseñar mejores políticas,	Si el modelo no es preciso, puede tomar decisiones equivocadas en cuanto a la promoción y desarrollo de los sitios turísticos.

		estrategias y planes de desarrollo turístico que impulsen la competitividad de los destinos.	Lo cual puede llevar a una perdida monetaria o mala imagen.
Asociación Hotelera y Turística de Colombia (COTELCO)	Socio / Beneficiario (seguidor)	Identificar las características que hacen atractivos a los sitios turísticos más destacados. Implementar estrategias para aumentar la calidad y satisfacción de los visitantes en los sitios menos recomendados.	En caso de que el modelo no funcione es dinero mal invertido y se pueden identificar de manera incorrecta los lugares que no cumplan con las expectativas de los turistas.
Universidad de los andes	Proveedor	Garantiza el cumplimiento de estándares de calidad de los modelos desarrollados, que incluye métricas de los datos utilizados y una explicación para la empresa.	Manejo incorrecto de los datos que lleve a la violación de la privacidad de los datos
Pequeños Hoteles en Municipios de Colombia	Socio / Beneficiario	Recibe un modelo con los datos clasificados al tiempo que le permite pensar en soluciones para varios problemas en simultaneo para problemas parecidos.	Recibir un modelo el cual no sea confiable o no este bien realizado, puede llevar a un aumento de afluencia lo cual al implementar la solución se encontrará con una peor solución.
Turistas Locales e Internacionales	Usuario Final	Recibir recomendaciones	Si el modelo clasifica

		de sitios turísticos que cumplan con sus expectativas y preferencias.	erróneamente los sitios, los turistas pueden tener experiencias negativas al visitar destinos que no cumplen con lo esperado.

(aclaración: se decidió poner al turista como el usuario final ya que, aunque las las diferentes organizaciones los datos y sean un usuario en general, el turista es el que va a tener las recomendaciones del modelo, por lo cual es el usuario final de todo. Cabe aclarar que el usuario de los datos y las métricas son las organizaciones)

(8%) Trabajo en equipo

- **Líder de proyecto: Medina Martínez, Ana Sofia**
- **Líder de negocio: Castellanos Bonilla, Juan Diego**
- **Líder de datos: Pérez Castilla, Carlos Mario**
- **Líder de analítica: Medina Martínez, Ana Sofia**

Reuniones:

- **Reunión de lanzamiento y planeación:** se llevó a cabo el lunes 18 de marzo, en el cual se definió el problema y se buscaron algoritmos para la solución del problema.
- **Reunión de ideación:** Se llevo a cabo el viernes 22 de marzo, en la cual se realizó la limpieza de datos de manera grupal, con el fin de que todos lo entendieran y se realizó la introducción del documento, para definir la organización, rol y beneficiario de la solución.
- **Reuniones de seguimiento:** se realizaron 3 reuniones de seguimiento. La primera el día 25 de marzo para saber que todos revisaron los datos y entendieron el problema. El día 29 de octubre para revisar la implementación de los algoritmos y avance del proyecto. La última reunión se hizo el día 2 de abril con el fin de seleccionar y explorar que puede afectar la precisión de cada algoritmo.
- **Reunión de finalización:** Una vez realizada la reunión de terminación del proyecto se verifico el trabajo y se decidió una fecha extra para explicarse entre todos cada uno de los algoritmos y todos conocer lo que se realizó en el trabajo. De igual manera para mejorar

se acordó no depender del trabajo de otras personas para continuar el trabajo individual.